DOCUMENT RESUME

ED 440 126 TM 030 738

AUTHOR Ediger, Marlow
TITLE Making Comparisons among Schools: The Report Card.
PUB DATE 2000-02-21
NOTE 11p.
PUB TYPE Opinion Papers (120)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; Elementary Secondary Education; *Institutional Characteristics; *Report Cards; *Schools; Test Results; Test Use

ABSTRACT

When schools are compared, there are variables that need to be considered because the playing field is not level. This paper discusses concepts and generalizations that must be understood to interpret test results and school report cards properly. There are many limitations in using test results to compare schools, particularly since students may be tested once a year at the most. The "one size fits all" approach of standardized testing fails to give a true picture of student learning and school effectiveness. Report cards for schools frequently omit important information such as expenditures, resources, and the condition of the school facilities. There are many problems in comparing longitudinal groups of students that may result from characteristics of the tests taken in different years or from differences in student characteristics. New assessment techniques, including portfolios, offer promise for providing data about schools, but care must be taken in comparisons based on qualitative data as well as comparisons based on quantitative data. (Contains 10 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 440 126

Making Comparisons Among Schools: The Report Card

Marlow Ediger

TM030738

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Marlow Ediger

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

MAKING COMPARISONS AMONG SCHOOLS-- THE REPORT CARD

Americans have become somewhat obsessed with documenting student achievement. Many tests are given annually to public school students in attempting to determine what they have learned. Teachers and administrators need to become highly familiar with how these comparisons are made. They also need to learn much about the concepts of measurement and evaluation. Student achievement is measured to ascertain a numerical result such as a percentile rating. The percentile rating for a student then needs to be evaluated in terms of its worth for that particular student. Selected students should and do achieve higher than do others due to increased abilities possessed.

When schools are compared with one another, there are salient variables that need to be considered since the playing field is not level by any means. Suburban students achieve at a much higher rate as compared to urban and rural school students. Which are selected concepts and generalizations that need to be understood to interpret test results and report cards properly?

Variables to Consider in Testing and Comparing

There are numerous variables that enter in to the interpretation of one student as well as entire school achievement of learners. When standardized tests are used to ascertain achievement, the following ideas are salient to understand:

1. they have no accompanying objectives for teachers to use in teaching, thus minimizing validity in in terms of content taught as compared to what is being tested.
2. they have built in features to spread students out on test results, from the 99th to the first percentile.
3. they standardize test taking to provide each student with the same directions for test taking, the same amount of time allotted in test completion, and use of the same key to check answers for the test results.
4. they tend to be high on reliability, be it test/retest, split/half, or alternative forms.
5. they provide data on student achievement with a single numeral, such as percentiles, standard deviations, quartile deviations, and/or grade equivalents. No information is then generally provided on what a student missed specifically for diagnostic and remediation purposes (Ediger, 1994, 169-174).

If criterion referenced tests (CRTs) are used on the report card to make comparisons among schools within a state, the following ideas are

important for teachers and administrators to know:

1. they do have accompanying objectives for teachers to use in teaching and be accountable for; the test items are more valid as compared to standardized tests since they tend to be aligned with the stated objectives.
2. they may not have been pilot tested adequately for clarity of test items and for reliability, since the money is not there to do this as compared to commercial for profit standardized tests.
3. they also dwell on multiple choice test items, as do standardized tests; lower levels of cognition then tend to be measured.
4. they too are machine scored so that mass numbers of student's test results may be revealed on printouts.
5. they indicate student achievement with a single numeral such as a percentile (Ediger, 1999, ERIC).

With the use of standardized tests and CRTs, a one shot case is in evidence since students may be tested once a year at the most. This leaves out student assessment from the 180 day daily work per school year engaged in by learners. A single score, such as a percentile, is to "tell it all" about student achievement when using standardized tests. There is much lacking here on data pertaining to learner achievement, such as what results are there from each student pertaining to a single lesson taught on a specific day of the school year. It is not possible to diagnose and remedy student difficulties in learning with a single percentile that is provided from standardized and CRTs.

Second, students need to reveal optimal achievement during testing time only, since these results serve as measures for reporting report card results. One can perceive students feeling ill, not up to par, upset emotionally, tense, and anxious during the one shot time for testing. Whereas, during an entire school year, there is a better opportunity to notice student achievement from the daily school work accomplished.

Third, the "one size fits all" is in evidence from standardized and CRTs. The same test, time limits, and directions given for test taking, among other factors, are the same for all students whereas in every day class work clarifications may be provided to students as needed to assist more optimal continuous achievement and progress. Students differ from each other in many ways and individual differences need to be provided for.

Fourth, when making comparisons among school systems for a report card, there are too many variables that are omitted when test results are reported. Thus, the playing field is not level. Minorities will not do as well as students from suburbia. Why? The environment in the home and community are not favorable for learning for many minority students. Opportunities are lacking here for learning and achieving. It is

surprising that minority students do as well as they do! Suburbia and its wealth has many educational opportunities to offer its young people that low socioeconomic levels do not have. The author tends to think that both standardized, in particular, and CRTs measure socioeconomic levels rather than achievement in academic knowledge (See Ediger, 1995, ERIC).

Fifth, are academic learnings the only important factor for students? Not all, by any means, will benefit from the academics only. Students differ in interests and abilities. Certainly, each student should be exposed to career education. It is not a shame to become a carpenter, mechanic, plumber, or technician. The writer took three years of vocational agriculture in high school and obtained the rank of State Farmer in the Future Farmers of America (FFA) organization. He also received a scholarship to Kansas State Agricultural College (KSAC) now named Kansas Sate University, at Manhattan, Kansas. Career and vocational education are good and honorable. Why should the academics be held in higher esteem? I believe too much time is wasted in teaching if all students are to go the academic route since many of these will remember little about these academic learnings. The drop out rate may be higher too if all are to be taught the academics only with the hope that "equality" will be an end result. Perhaps, the concept of equality needs redefining since not all are going to be involved, by any means, with the academic at the future work place. There are essential learnings for all in the 3 r's, social studies, science, art, music, and physical education. Beyond that, educators need to think about what should be tested upon, e.g. should the curriculum go beyond the academic world? Then to, testing involves assessing verbal intelligence, such as reading test items largely. Multiple intelligences Theory stress the importance of students revealing in additional ways that which has been learned (See Gardner, 1993).

Report cards too frequently omit important information pertaining the following:

1. how much a school system spends on school supplies and teacher salaries.

2. the condition of the school building. The writer taught in a school building in which the roof leaked very badly. Buckets were set up in his classroom to catch the falling rain drops with the accompanying continuous unpleasant, annoying sounds of "splish, splash, spat, and splash." The writer then capitalized on the sounds by having students write poetry with alliteration. The writer also taught in a rural school where the water table went zilch in the morning, and the county superintendent of schools recommended keeping the school in session! That was a very bad recommendation indeed.

3. the involved heating system and its operations. In the same school building with the leaky roof, the heating system consisted of

steam radiator pipes. The banging noise of the pipes made me feel as if a "ghost" was in the building hitting these pipes leading to the different classrooms. The temperature certainly varied much from 90 degrees to forty degrees Fahrenheit on a cold day.

Air conditioning is needed for hot days in early fall and late spring as well as for summer school. A good summer school program should be available for all students who desire it to be in operation

3. the quality and number of library books in the school that are used to encourage student reading accomplishment.

4. available modern technology in the curriculum to provide for individual differences in the classroom.

5. adequate and high quality support personnel services, such as guidance counselors and school health nurses, as well as .social services to assist in obtaining desirable food, clothing, and shelter for needy individuals (See Ediger, 1998, 541-548).

Report cards then need to show more than test data of learners. Test data, such as numerical scores, may reveal little in terms of student achievement and progress. Thus, assistance based on diagnosis needs to be provided to help students achieve more optimally. Meeting physiological, safety, belonging, and esteem needs are vital for each (See Maslow, 1954). Otherwise, achievement of students will be at a lower level.

Group Scores

Scores on a report card may be given over a period of time, such as several years.Cohort scores may then be provided covering five sequential years. Thus, for example, fifth graders may be compared pertaining to the school years including 1994- 98 school years. But, these are not the same fifth graders being compared each school year. Each school year has a different set of fifth graders. It might be that for each school year the fifth grades differ <u>much</u> from each other.

The same fifth graders may be compared covering five sequential school years. The mean gains from the first to the fifth school year may then be compared to notice if the gains are significant in a longitudinal study. The cohort study may also be compared in mean gains from the first to the fifth school year, but each year of schooling there is a different set of fifth graders.These kinds of comparisons are called cross sectional studies.

Longitudinal studies have more worth as compared to cross sectional studies in that the same fifth graders are used for the five year period.

Second, if the means of an experimental group is compared with a control group, a random sampling procedure should be used for both groups. If this is not done, one of the two groups may be ahead initially

before the study is begun. If the two groups are not equal initially, analysis of covariance may be used to statistically equalize the two groups. The analysis of covariance is stressed as a statistical procedure to equalize the means initially of the experimental and the control group. If the two groups, the experimental with the new procedure and the control with the traditional method, do not start at the same place of mean achievement initially, the results may mean nothing.

Third, very frequently to be significant statistically, the end results between the experimental and the control groups need to be at the .05 level. Sometimes the hypothesis to be tested between the final means of the two groups is less significant than at the .05 level, such as .06 level. Does the study then means nothing since the results were not significant at the e.05 level? The reader of the research needs to study this and realize it was close to being significant at the .05 level. A judgment should then be made by the reader of the research to ascertain how important the results were.

Fourth, rank order scores may provide some difficulty in interpretation. If school systems on a report card are ranked from top to bottom, based on test score results, school A may be at the top, followed by school B, and then school C, and so on. But, what if the gaps among these three schools are so very small in terms of raw score points in school achievement based on standardized test results? School A may average a raw score of 85, school B 84, and school C 83. Suppose the Standard Error of measurement (SE meas) was two raw score points. Then school A's raw score could vary from 83 to 87, school B from 82 to 86, and school C from 81 to 85, due to error in the tests and in testing. It truly is difficult to say which of the three schools had the best average test results form students.

Fifth, tests used may have so many weaknesses, that when used to make comparisons among schools, may mean little or nothing. Validity data given in the eleventh edition of the Mental Measurement Yearbooks need to be studied in terms of testing and measurement quality for the test being used. If the standardized test is older than 1995, an earlier edition of the Mental Measurement Yearbook needs to be consulted.These yearbooks represent a tester's Bible and, no doubt, provide the best information possible pertaining to a critical review of each standardized test. Testing and measurement specialists provide these reviews. In addition to validity data, information on reliability, among other items, should also be evaluated as given in the Mental Measurement's Yearbook for the test used in doing research. Consumers of educational research data should be skeptical of how schools are rated on a report card. There are many variables that go into school achievement or lack thereof.

Sixth, writers in education may have their biases and agendas. The reader of research needs to be very skeptical of a writer who

advocates that only the following procedures and methods of instruction should be used:

a) heterogeneous grouping with no homogeneous grouping.

b) cooperative learning with no individual endeavors for students in class.

3) focus upon the academics in teaching only. There is much more to learning than subject matter only, such as ethics, character, and education for democracy as a way of life.

4) gender education focusing on female students only with complete omissions on assisting boys to also achieve as optimally as possible.

5) measurable evaluation results only to the exclusion of qualitative assessment (See Andrade, 2000).

From Quantitative to Qualitative Assessment

Quantitative results provide numerical data only, from student's tested achievement. To remedy deficiencies here, qualitative procedures have come into the offing. Portfolio use is a good example. Portfolio results shifts philosophy of assessment from measurement to constructivism. Constructivism emphasizes assessing learner progress within an ongoing learning experience. It stresses continuous evaluation in ongoing lessons and units of study. The classroom teacher together with the involved student(s) might then appraise the latter's achievement Assistance might be provided on the spot to guide students to achieve, grow, and develop. Objectives of instruction provide a benchmark for what is to be taught. The accomplishments are not haphazard but are based upon the objectives to be stressed in teaching and learning. Validity should be high here if the products/process of instruction match with the objectives of instruction. There still is room here to incorporate student objectives and aims.

The portfolio stresses heavy input from the student as to what should comprise the final product here. The contents in a portfolio are purposeful in that they indicate what has been achieved by a student. They represent a random sampling of accomplishments by the learner covering a specific period of time. The contents of a portfolio indicate what a student has achieved on a daily basis. Care, however, much be in the offing to make the contents representative, and not become too voluminous. What might go into a portfolio for a student?

1. written work such as outlines, essays, reports, summaries, and conclusions, among others.
2. art products as they relate to ongoing units of study.
3. cassettes of oral communication.
4. snapshots of projects too large to place in a portfolio.
5. a video of committee work showing efforts of the involved

student.

6. creative work revealing prose and poetry.
7. statements of self evaluation.
8. journal entries that relate to daily experiences in the classroom.
9. goal setting by the student to achieve at voluntary tasks such as at enrichment centers, learning stations, and library work.
10. homework completed by the learner (Ediger, 2000, Chapter Ten).

The contents of a portfolio then do not permit specific numerals to be given for achievement results as is true of standardized and criterion referenced tests. Rather a qualitative approach is used in assessment. Generally, several professional teachers are recommended to assess a portfolio. Rubrics can be used to make the results for a portfolio evaluation more objective. Usually, a four or five point scale is used to show the rating of a portfolio. If four levels are used in rating a portfolio, each of the four needs to specify what is expected for the highest rating, and other ratings, to be given. When going by the specifications for each of the four levels a student may receive, the ratings should become more objective. Thus, increased reliability is in the offing when the different rates agree about the quality of a portfolio. However, portfolios do represent a qualitative rather than a quantitative procedure of assessment.

As is true of all assessment procedures of student achievement, portfolios have their weaknesses. Among others, the following need to be looked at with improved procedures being in the offing:

1. reliability in portfolio appraisal probably will be somewhat low. Why? Scorers of the portfolio will not agree upon results for any one portfolio. Interscorer ratings then will vary from one rater to another on the same portfolio. A remedy here might be increased inservice education for raters so that more agreement is possible on how to assess a portfolio.
2. much time can be spent on portfolio assessment. One portfolio evaluated by two to three teachers will take up a considerable amount of time. If twenty portfolios are appraised by two teachers, the time given here might well be great. This may take time away from teaching and learning situations. Machine scoring of portfolios is not possible.
3. rubrics used for assessment of portfolios may lack clear descriptions as to which portfolios should have ratings of one through four or five. It is very difficult to used descriptive statements and use these to assess portfolios. The descriptive statements for each category of the four to five point scale to assess portfolios should be precise and clear. Increased objectivity in assessment should be an end result.
4. items for a portfolio are difficult to choose in order that a random sampling of a student's work is in evidence.
5. many rubrics will need to be developed to appraise contents in

a portfolio. Why? A rubric for an essay written by the involved learner cannot be used to assess the following entries: a poem, an oral report, a narrative account, a construction item, an art project, a dramatics experience, and a discussion setting. These activities are common to use in unit teaching in any curriculum area and will need separate rubrics for evaluation purposes (See Salvia and Ysseldyke, Chapter Twelve).

How "objective" are standardized tests? They also lack objectivity in the following ways:

1. test writers could choose other items for the test than those selected.
2. human beings write the test items. The human factor does not make for objectivity.

Objectivity for standardized tests enters in with the following:

1. directions for administering the test, after these have been written by human beings (the test writers), are the same for all test takers.
2. the scoring key, once agreed upon, is used in scoring all test results.
3. time limits for taking the test, once agreed upon by the test writers, is the same for all who take the test.

Conclusion

There are numerous areas of disagreement on how students should be assessed to indicate achievement. the following are selected issues in the disagreement:

1. quantitative versus quantitative methods.
2. standardized "one size fits all" versus providing for individual differences such as a single student's portfolio.
3. annual reports on a report card or an individual's test results provided numerically as compared to ongoing assessments of a learner's progress in the classroom.
4. outsiders involved in determining what should be tested upon, such as writers of standardized and CRTs, versus contextual assessment in the local classroom, on a continuum.
5. sporadic assessment such as once a year, versus ongoing evaluation of a student's progress.

References

Andrade, Heidi Goodrich (2000), "Using Rubrics to Promote Thinking and Learning, Educational Leadership, 57 (5), 13-19.

Ediger, Marlow (1994), "Measurement and Evaluation," Studies in Educational Evaluation, 20 (2), 169- 174.

Ediger, Marlow (1999), "Issues in Appraising Achievement," Resources in Education, Educational Resources Information Center (ERIC).

Ediger, Marlow (1995), "To Every Action, There is an Opposite and Equal Reaction, Resources in Education, Educational Resources information Center (ERIC) # ED 386319.

Ediger, Marlow (1998), "Change and the School Administrator," Education, 118 (4), 541-548.

Ediger, Marlow (2000), Teaching Science in the Elementary School. Kirksville, Missouri: Simpson Publishing Company, Chapter Ten.

Gardner, Howard (1993), Multiple Intelligences: Theory Into Practice. New York: Basic Books.

Maslow, Abraham (1954), Motivation and Personality. New York: Harper and Row.

Mental Measurements Yearbook (1995), eleventh edition. New York: Gryphon Press.

Salvia, John, and James E. Ysseldyke (1995), Assessment, Sixth Edition. Boston: Houghton-Mifflin Company, Chapter Twelve.